HDAT9700: Presenting and summarising model results

Dr Md Shajedur Rahman Shawon
CBDRH, UNSW Sydney

Learning outcomes

- Evaluate characteristics of good figures and tables
- Understand how to report statistical significance by interpreting P-values and confidence intervals
- Communicate the estimates for measures of association
- Identify key guidelines and checklists for reporting of statistics

Housekeeping

- No formal assignment for this chapter
- 10% marks for your final report will be based on presentation
- Pre-readings:
  - Kenneth Rothman Six Persistent Research Misconceptions J Gen Intern Med. 2014 Jul; 29(7): 1060–1064
  - Professor Gary King's video on Presenting and Interpreting Statistical Results [Available on YouTube]

Health data scientists

Our main tasks:

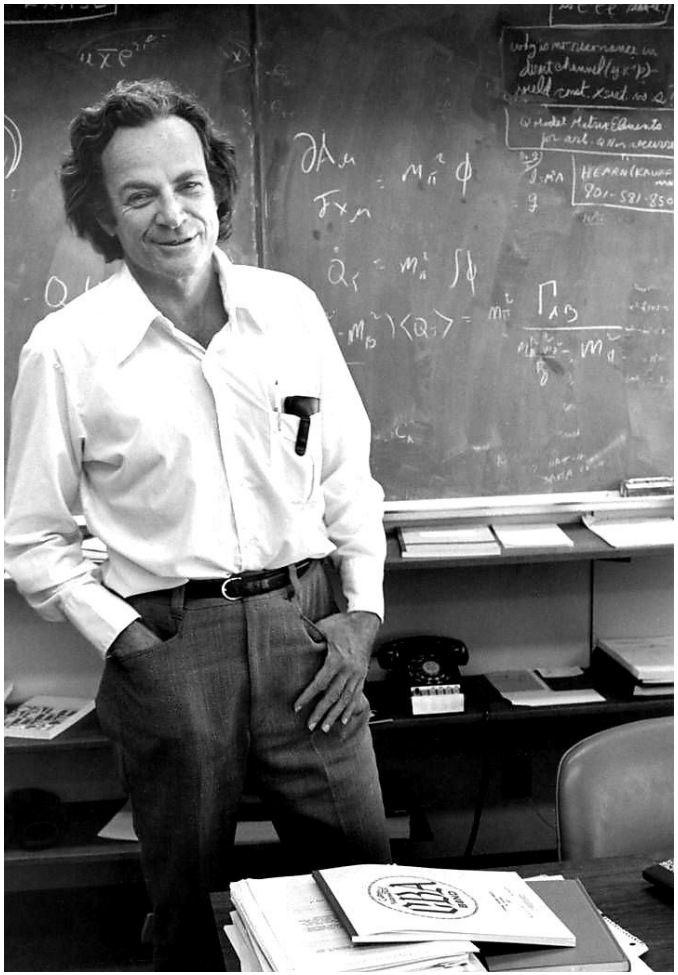
- Data cleaning and management
- Data analysis
- Finding the perfect (!) model that explains our data

Then why bother about data presentation?

Why focusing on presenting and summarising results?

- Readers of our research are humans (busy + inattentive + lazy to some extent)
- First chance to create impressions of your work
- The probability of your research having an impact
- Meeting scientific standards
- More importantly, scientific communication is an important part of science

My favourite science communicator



UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

What are the ways to present results?

Model results can be presented as

Table

Figure

Text

Tables vs. Figures

- Tables and figures are the backbones of almost all scientific documents.
- Readers often look at them first, even before reading any text

Figures are helpful because

- Figures have a strong visual impact
- Figures are used to show trends and patterns in data
- Figures can be used to highlight the most important results

Tables can be helpful because

- Tables can give precise values
- We can display a lot of information (many variables with many values) in a table

How to make a good table?

Table 1. Demographic and Clinical Characteristics of the Patients at Baseline.*

Characteristic	Early Rhythm Control (N=1395)	Usual Care (N=1394)
Age — yr	70.2±8.4	70.4±8.2
Female sex — no. (%)	645 (46.2)	648 (46.5)
Body-mass index†	29.2±5.4	29.3±5.4
Type of atrial fibrillation — no./total no. (%)		
First episode	528/1391 (38.0)	520/1394 (37.3)
Paroxysmal	501/1391 (36.0)	493/1394 (35.4)
Persistent	362/1391 (26.0)	381/1394 (27.3)
Sinus rhythm at baseline — no./total no. (%)	762/1389 (54.9)	743/1393 (53.3)
Median days since atrial fibrillation diagnosis (IQR)‡	36.0 (6.0–114.0)	36.0 (6.0–112.0)
Absence of atrial fibrillation symptoms — no./total no. (%)§	395/1305 (30.3)	406/1328 (30.6)
Previous cardioversion — no./total no. (%)	546/1364 (40.0)	543/1389 (39.1)
Concomitant cardiovascular conditions		
Previous stroke or transient ischemic attack — no. (%)	175 (12.5)	153 (11.0)
At least mild cognitive impairment — no./total no. (%)¶	582/1326 (43.9)	584/1341 (43.5)
Arterial hypertension — no. (%)	1230 (88.2)	1220 (87.5)
Blood pressure — mm Hg‖		
Systolic	136.5±19.4	137.5±19.3
Diastolic	80.9±12.1	81.3±12.0
Stable heart failure — no. (%)**	396 (28.4)	402 (28.8)
CHA_2DS_2-VASc score††	3.4±1.3	3.3±1.3
Valvular heart disease — no./total no. (%)	609/1389 (43.8)	642/1391 (46.2)
Chronic kidney disease of MDRD stage 3 or 4 — no. (%)‡‡	172 (12.3)	179 (12.8)
Medication at discharge — no./total no. (%)§§		
Oral anticoagulation with NOAC or VKA	1267/1389 (91.2)	1250/1393 (89.7)
Digoxin or digitoxin	46/1389 (3.3)	85/1393 (6.1)
Beta-blocker	1058/1389 (76.2)	1191/1393 (85.5)
ACE inhibitors or angiotensin II receptor blocker	953/1389 (68.6)	979/1393 (70.3)
Mineralocorticoid-receptor antagonist	90/1389 (6.5)	92/1393 (6.6)
Diuretic	559/1389 (40.2)	561/1393 (40.3)
Statin	628/1389 (45.2)	568/1393 (40.8)
Platelet inhibitor	229/1389 (16.5)	226/1393 (16.2)

* Plus–minus values are means ±SD. Definitions of clinical measures are provided in Table S1. ACE denotes angiotensin-converting enzyme, IQR interquartile range, NOAC non–vitamin K antagonist oral anticoagulant, and VKA vitamin K antagonist.
† The body-mass index is the weight in kilograms divided by the square of the height in meters. Data were missing for 7 patients assigned to early rhythm control and for 6 patients assigned to usual care.
‡ Data on median days since atrial fibrillation diagnosis were missing for 2 patients assigned to early rhythm control and for 1 patient assigned to usual care.
§ The absence of symptoms was defined as a European Heart Rhythm Association (EHRA) score of I. The EHRA score categorizes symptoms related to atrial fibrillation into four classes from I (asymptomatic) to IV (severe symptoms at rest).
¶ At least mild cognitive impairment was defined as a Montreal Cognitive Assessment (MoCA) score of less than 26. The MoCA score provides an overall assessment of cognitive function. Scores range from 0 to 30, with lower scores indicating worse cognitive function.
‖ Data on blood pressure were missing for 9 patients assigned to early rhythm control and 4 patients assigned to usual care.
** Stable heart failure was defined as New York Heart Association stage II or a left ventricular ejection fraction of less than 50%.
†† CHA_2DS_2-VASc scores (an assessment of the risk of stroke among patients with atrial fibrillation) range from 0 to 9, with higher scores indicating a higher risk of stroke.
‡‡ A Modification of Diet in Renal Disease (MDRD) stage of 3 or 4 indicates a glomerular filtration rate of 15 to 59 ml per minute per 1.73 m^2 of body-surface area.
§§ Because of the high proportion of patients with atrial fibrillation that was first diagnosed at enrollment, important therapies were initiated between enrollment and discharge from the baseline visit. Therefore, medication at discharge from the baseline visit is shown.

Source: P Kirchhof et al. N Engl J Med 2020.

Points to be noted:

- Title
- Alignment
- Consistency
- No empty cells
- Use of headings
- Variable definitions
- Footnotes
- Abbreviations

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Can you tell the difference?

Factors	No infection	Infection
Age in years, mean (SD)	69.8 (11.2)	71.9 (12.4)
Female sex, n (%)	1245 (45)	321 (48)
Previous infection, n (%)	205 (7.2)	32 (10)

Factors	No infection	Infection
Age in years, mean (SD)	69.8 (11.2)	71.9 (12.4)
Female sex, n (%)	1245 (45)	321 (48)
Previous infection, n (%)	205 (7.2)	32 (10)

Can you tell the difference?

Factors	No infection	Infection
Age in years, mean (SD)	69.8 (11.2)	71.9 (12.4)
Female sex, n (%)	1245 (45)	321 (48)
Previous infection, n (%)	205 (7.2)	32 (10)

Factors	No infection	Infection
Age in years, mean (SD) Female sex, n (%) Previous infection, n (%)	69.8 (11.2) 1245 (45) 205 (7.2)	71.9 (12.4) 321 (48) 32 (10)

Avoid copy + pasting numbers

Instead use different packages to directly export your results to Excel or Word.

Advantages:

- Saves time and effort
- Reduces error
- Facilitates reproducible research

How to make a good figure?

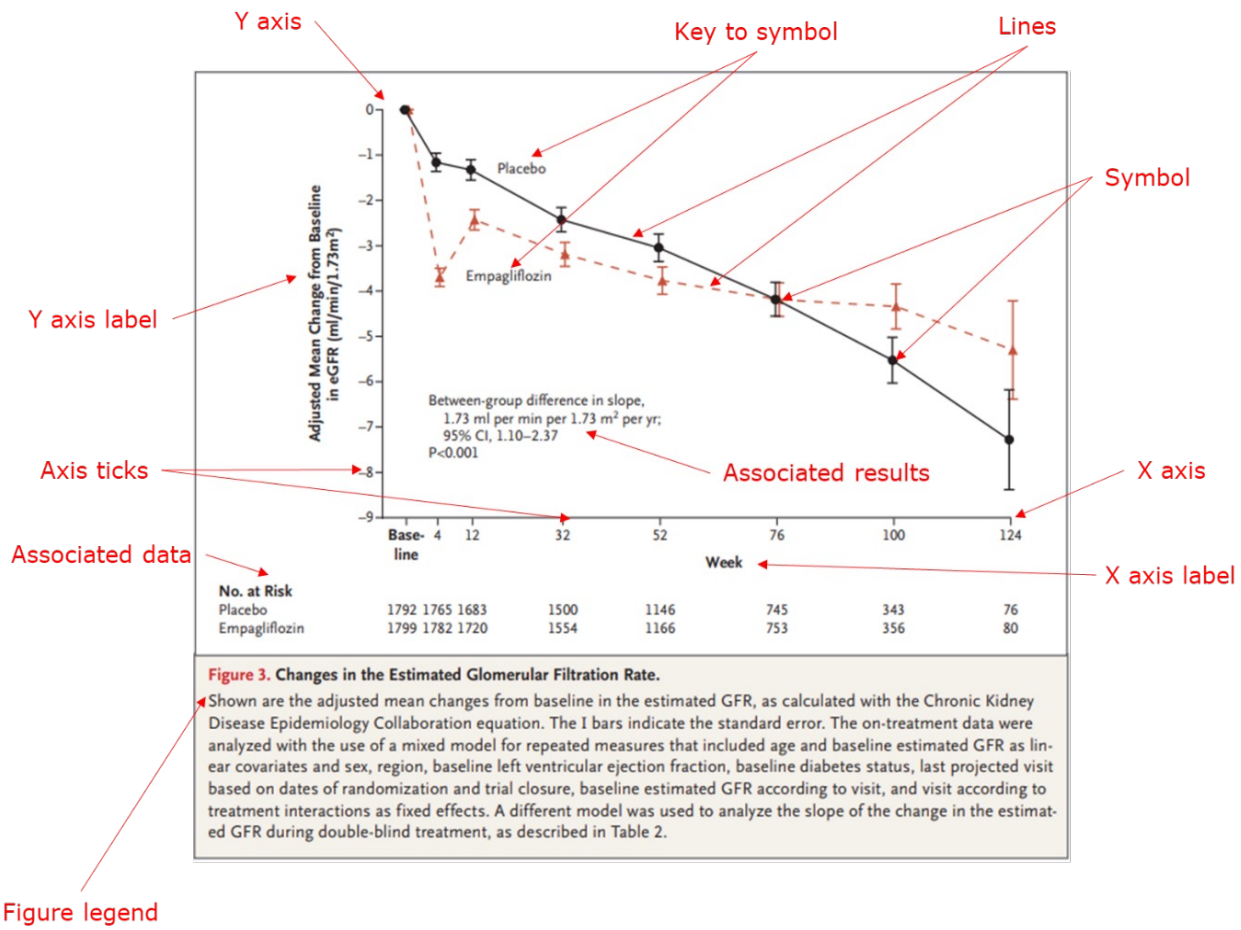


Figure 3. Changes in the Estimated Glomerular Filtration Rate.

Shown are the adjusted mean changes from baseline in the estimated GFR, as calculated with the Chronic Kidney Disease Epidemiology Collaboration equation. The I bars indicate the standard error. The on-treatment data were analyzed with the use of a mixed model for repeated measures that included age and baseline estimated GFR as linear covariates and sex, region, baseline left ventricular ejection fraction, baseline diabetes status, last projected visit based on dates of randomization and trial closure, baseline estimated GFR according to visit, and visit according to treatment interactions as fixed effects. A different model was used to analyze the slope of the change in the estimated GFR during double-blind treatment, as described in Table 2.

A figure legend may contain:

- A brief title
- Variable definitions
- Definitions of symbols or lines
- Explanation of panels (A, B, C, D)
- Statistical information (statistical model specification, tests used, p-values)
- Brief description of the key findings that are seen in the figure

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Message trumps beauty

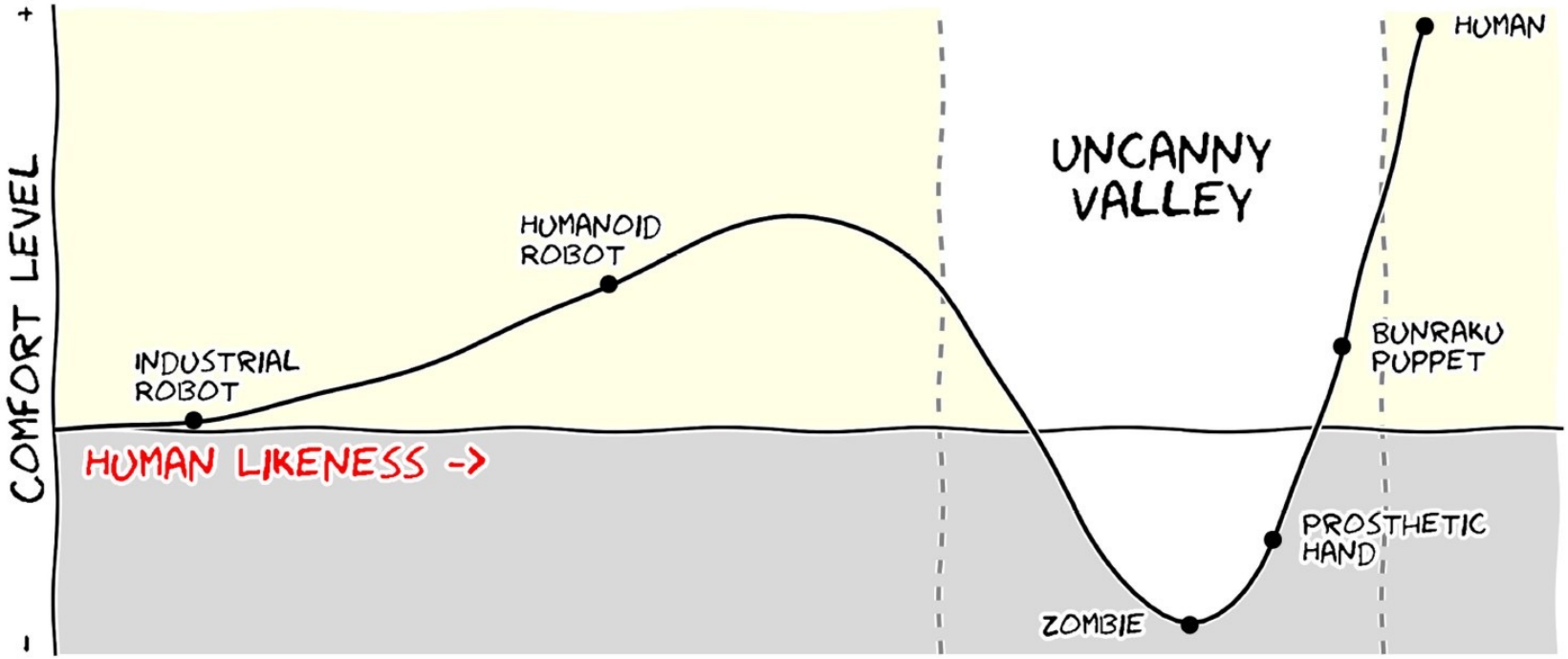


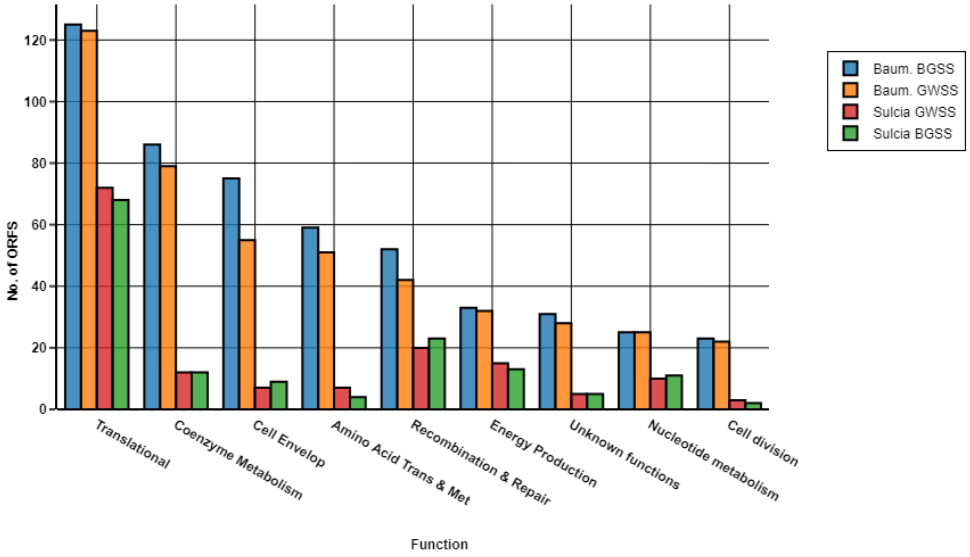
Figure 8. Message trumps beauty. This figure is an extreme case where the message is particularly clear even if the aesthetic of the figure is questionable. The uncanny valley is a well-known hypothesis in the field of robotics that correlates our comfort level with the human-likeness of a robot. To express this hypothetical nature, hypothetical data were used ($y = x^2 - 5e^{-5(x-2)^2}$) and the figure was given a sketched look (xkcd filter on matplotlib) associated with a cartoonish font that enhances the overall effect. Tick labels were also removed since only the overall shape of the curve matters. Using a sketch style conveys to the viewer that the data is approximate, and that it is the higher-level concepts rather than low-level details that are important [10].
doi:10.1371/journal.pcbi.1003833.g008

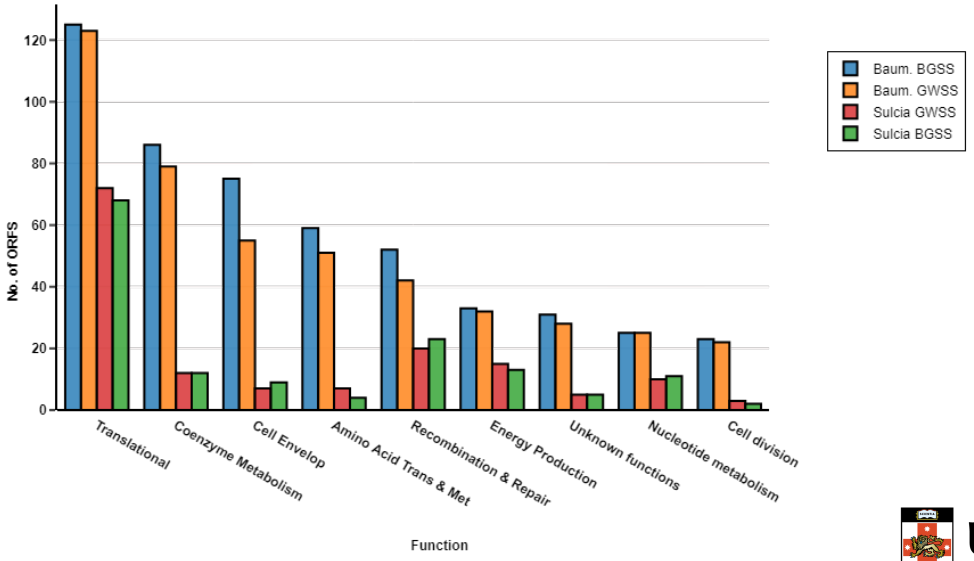
Avoid Chartjunk

“Perfection is achieved not when there is nothing more to add, but when there is nothing left to take away.”

— Antoine de Saint-Exupery.

Which graph looks better?





UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

For more tips

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Better Figures

Nicolas P. Rougier[1,2,3]*, Michael Droettboom[4], Philip E. Bourne[5]

1 INRIA Bordeaux Sud-Ouest, Talence, France, **2** LaBRI, UMR 5800 CNRS, Talence, France, **3** Institute of Neurodegenerative Diseases, UMR 5293 CNRS, Bordeaux, France, **4** Space Telescope Science Institute, Baltimore, Maryland, United States of America, **5** Office of the Director, The National Institutes of Health, Bethesda, Maryland, United States of America

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

DPI requirements by journals

- DPI stands for "dots-per-inch"
- Many journals ask for figures with having at least 300 dpi
- How to check the dpi of an image?

GIMP

- Open software
- GIMP is the GNU Image Manipulation Program. It is an application for such tasks as photo retouching, image composition, and image authoring.
- If you need to quickly retouch an image or add some legends or labels, GIMP is the perfect tool.

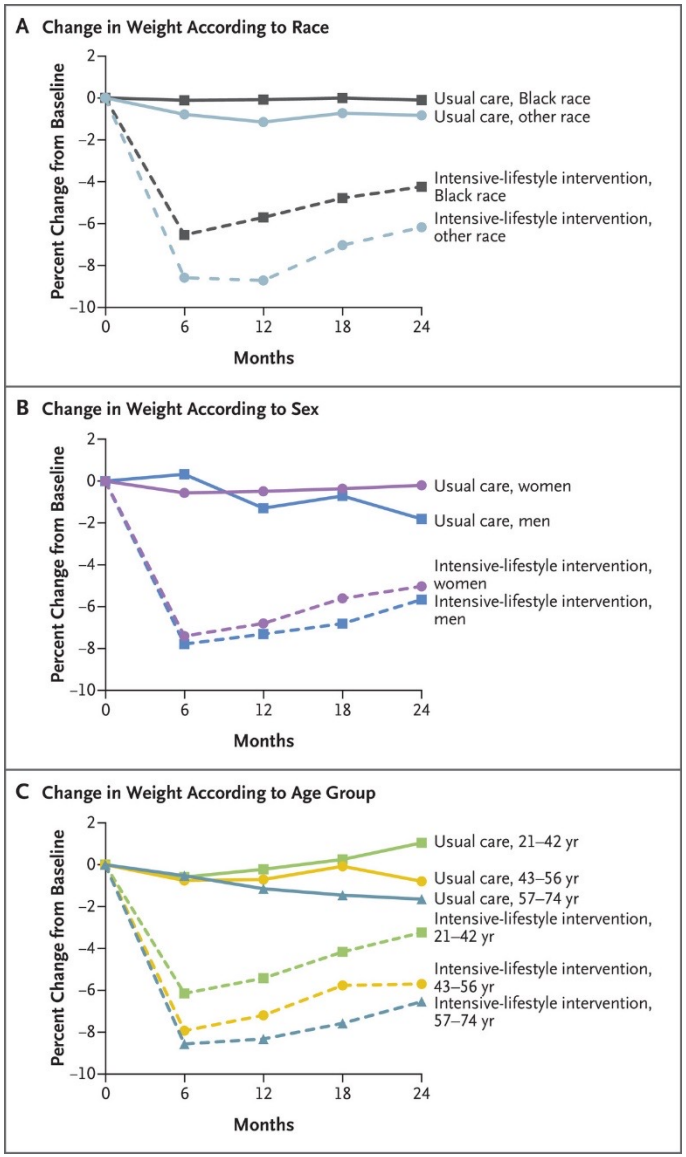
Multi-panel plots

- Combining multiple plots or add objects (e.g., image or text) to a single plot

Utilities:

- Reduce the number of total figures
- Convenience to the readers
- Can highlight the comparisons

- The legend of a multi-panel figure must identify each graph
- When mentioned in the text, specific graph needs to be cited [e.g. Figure 2 (A)]



Source: PT Katzmarzyk et al. N Engl J Med

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Creating tables and figures in R

- R packages with some exercises are given on the course website
- The rules for a good figure or table are independent of the package or statistical programs
- Do not trust the defaults
- Always refer to the journal guidelines

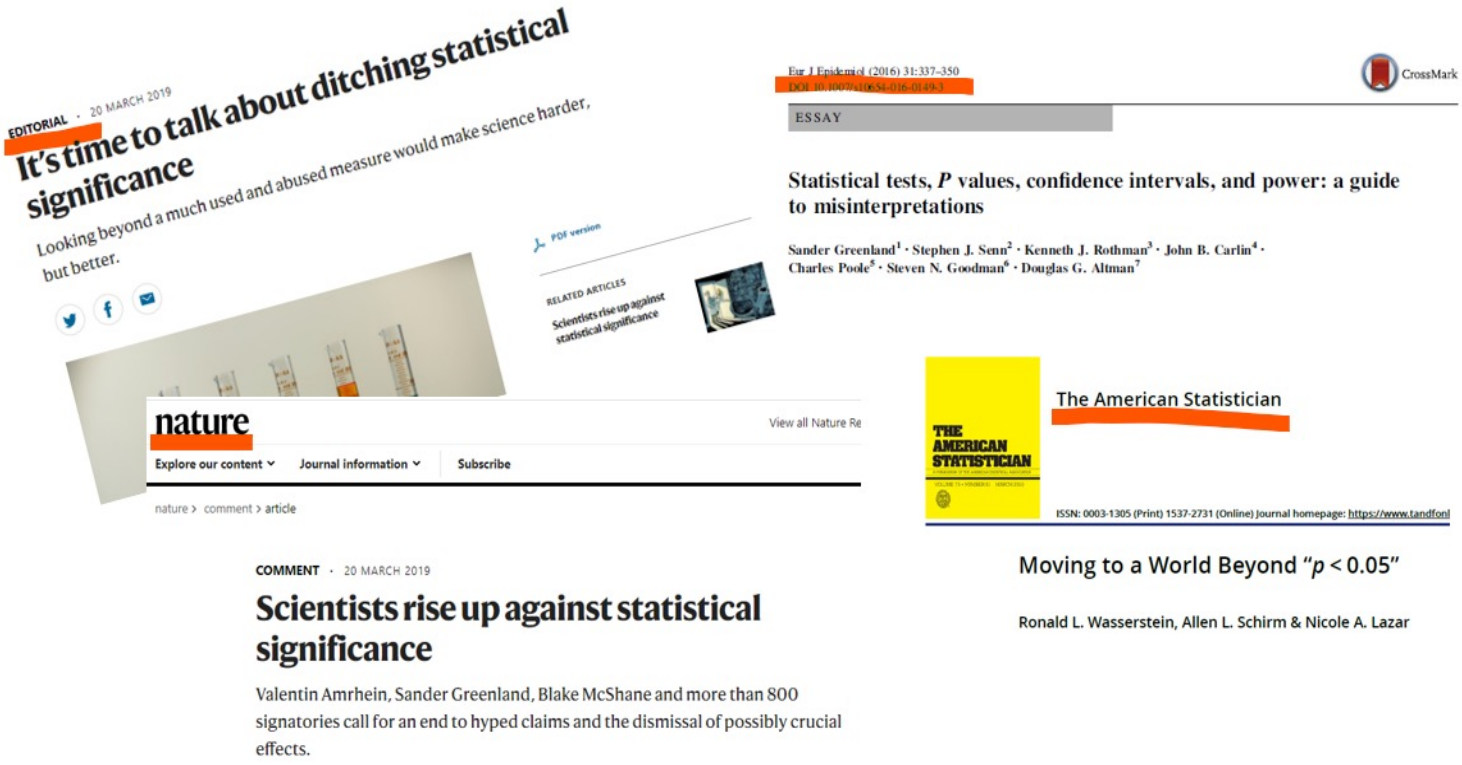
What are the common measures we get from statistical models?

Common measures we get from statistical models

- β – coefficient
- P-value
- 95% confidence intervals
- Measures of association:
  - Odds ratio
  - Hazard ratio
  - Rate ratio

- *Correct interpretation of them is key to the effective presentation of your results.*

Reporting statistical inference



"When a measure becomes a target, it ceases to be a good measure."

-Goodhart's law

UNSW SYDNEY

CENTRE FOR BIG DATA RESEARCH IN HEALTH

Why are the p-values different?

Study 1

- Male: 68.5 kg
- Female: 66.0 kg
- Difference: 2.5 kg
- P-value: 0.10

Study 2

- Male: 67.2 kg
- Female: 62.0 kg
- Difference: 5.2 kg
- P-value: 0.02

Study 3

- Male: 65.2 kg
- Female: 62.7 kg
- Difference: 2.5 kg
- P-value: 0.02

Why are the p-values different?

Study 1

- Male: 68.5 kg
- Female: 66.0 kg
- Difference: 2.5 kg
- P-value: 0.10
- Sample size: 240

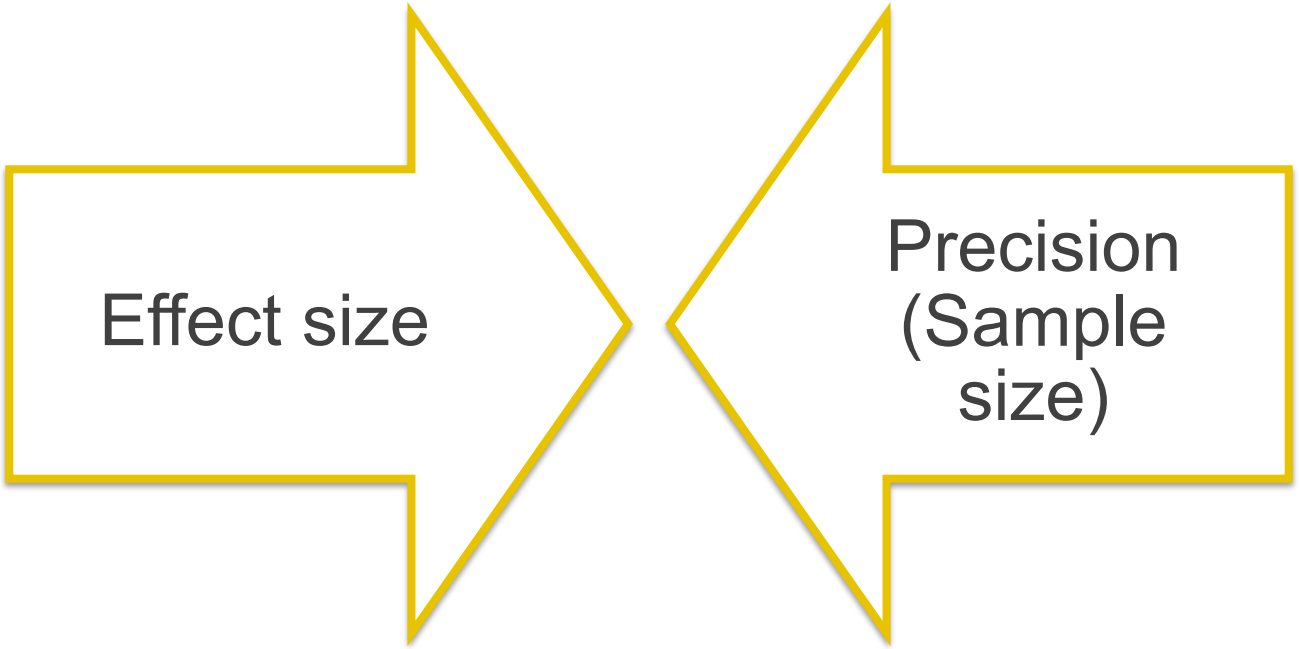
Study 2

- Male: 67.2 kg
- Female: 62.0 kg
- Difference: 5.2 kg
- P-value: 0.02
- Sample size: 240

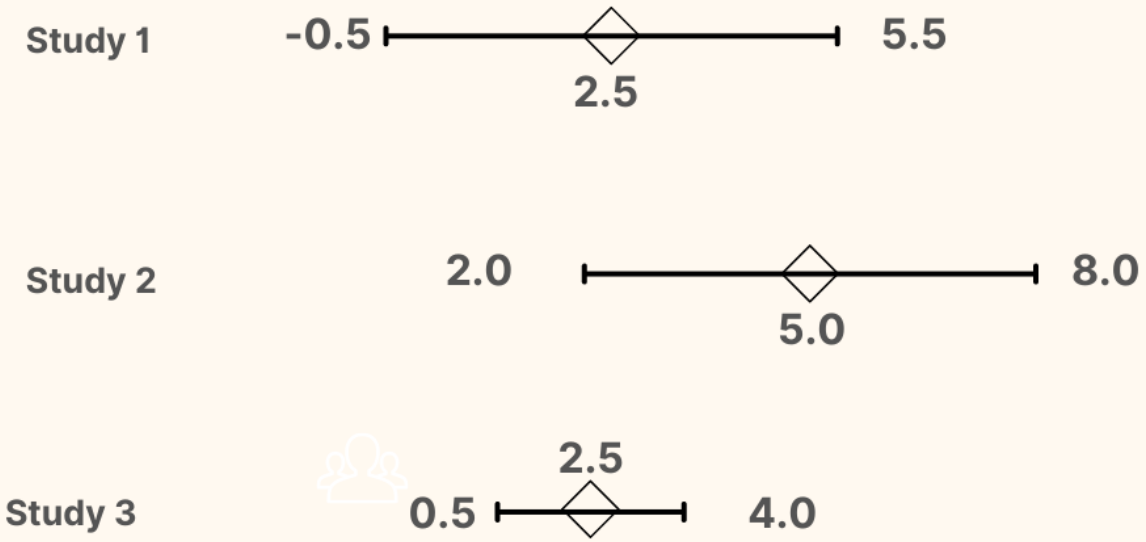
Study 3

- Male: 65.2 kg
- Female: 62.7 kg
- Difference: 2.5 kg
- P-value: 0.02
- Sample size: 980

P-value depends on



Confidence intervals



A shift from significance testing to estimation

Five Don'ts:

- Don't base your conclusions solely on whether an association or effect was found to be "statistically significant"
- Don't believe that an association or effect exists just because it was statistically significant.
- Don't believe that an association or effect is absent just because it was not statistically significant.
- Don't believe that your p-value gives the probability that chance alone produced the observed association or effect or the probability that your test hypothesis is true.
- Don't conclude anything about scientific or practical importance based on statistical significance (or lack thereof).

Two Dos:

- Do use confidence intervals as quantitative measures indicating magnitude of effect size and degree of precision
- Do accept uncertainty: P-values, confidence intervals, and other statistical measures are all uncertain.

Interpreting measures of association

Measures of association:

In relative scale:

- Relative risk
- Incidence rate ratio
- Hazard ratio
- Odds ratio

In absolute scale:

- Risk difference

Hypothetical example

Table 1: Hypothetical results of an 8-week randomised controlled trial of venlafaxine (n = 40) vs placebo (n = 36)

	Developed sexual dysfunction	Did not develop sexual dysfunction	Total
Received venlafaxine	8	32	40
Received placebo	3	33	36
Total	11	65	76

- The relative risk of sexual dysfunction with venlafaxine vs. placebo is 2.40
- The odds ratio of sexual dysfunction with venlafaxine vs. placebo is 2.75

Interpretations of relative risk

- Venlafaxine, relative to placebo, is associated with a 2.4-fold risk of sexual dysfunction.
- Venlafaxine is associated with a more than doubled risk of sexual dysfunction.
- The risk of sexual dysfunction with venlafaxine is 2.4 times that with placebo.
- The risk of sexual dysfunction with venlafaxine is 240% that with placebo.
- Compared with placebo, venlafaxine is associated with a 1.4-fold increased risk of sexual dysfunction.
- Compared with placebo, venlafaxine is associated with a 140% increased risk of sexual dysfunction.

Interpretations of odds ratio

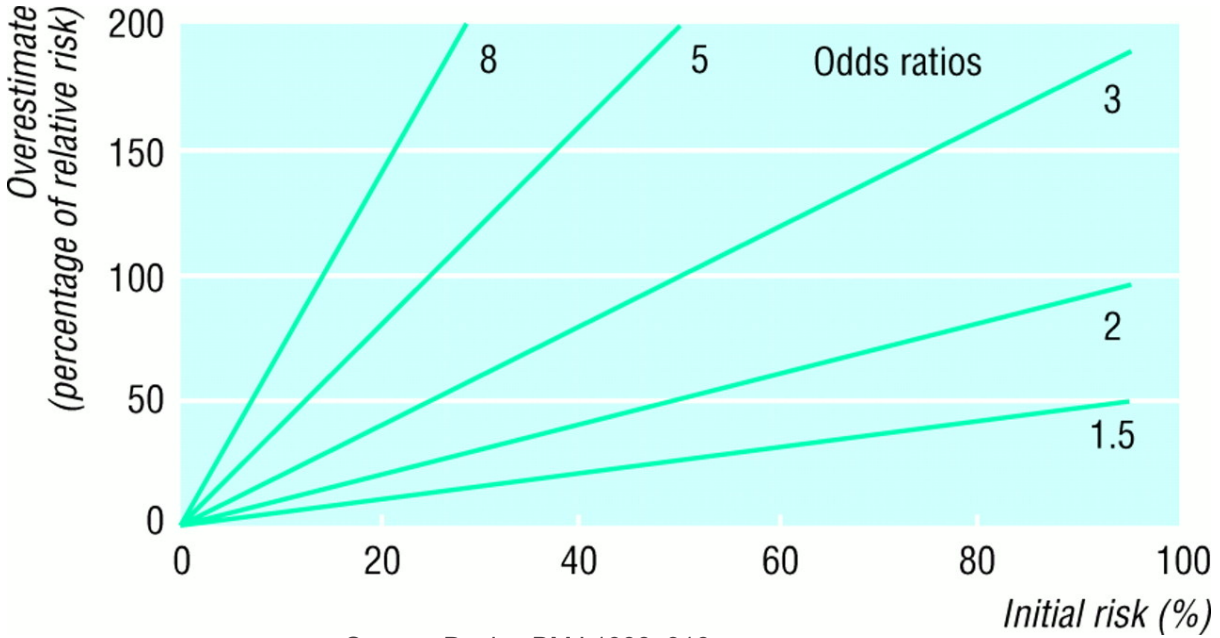
The communications for odds ratios are not as intuitive as relative risks, particularly because odds are not an easy-to-understand concept like probabilities.

- The odds of developing sexual dysfunction with venlafaxine (relative to placebo) are 2.75 to 1.
- The odds of developing sexual dysfunction with venlafaxine are 2.75 as large compared with placebo.
- There is a 175% increase in odds of developing sexual dysfunction with venlafaxine when compared with odds of developing sexual dysfunction with placebo.

Translation of an odds ratio of X to an “X-fold risk,” is seen commonly in published papers. This is wrong because odds ratios are based on a different underlying concept than relative risks.

Are odds ratio and relative risk equivalent?

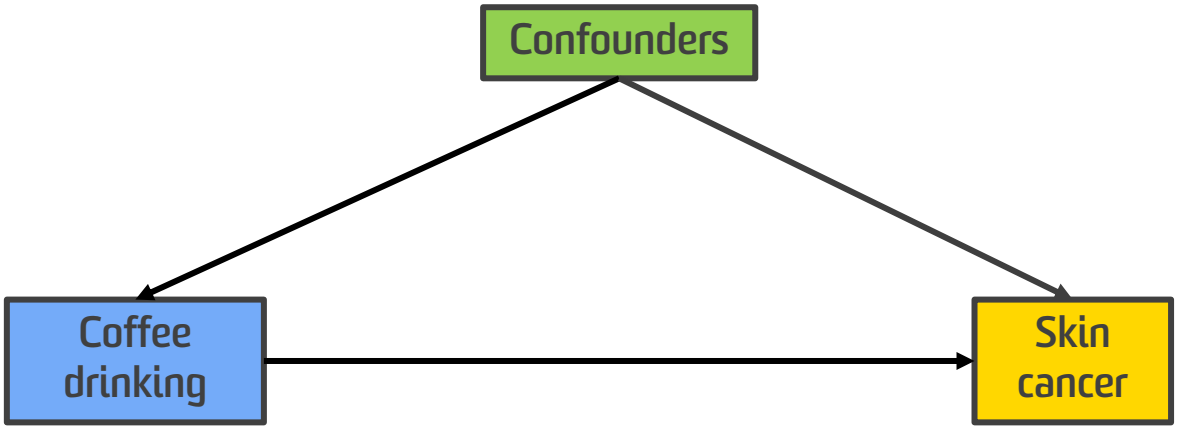
- Rare disease assumption
- When an event occurs more commonly, the odds ratios always overstate any effect size whereas relative risks remain constant.
- It is also important to remember that because of the mathematical behaviour of the odds ratio, estimates of the odds ratio may vary among studies of differing designs, even if the actual risk ratio is constant.



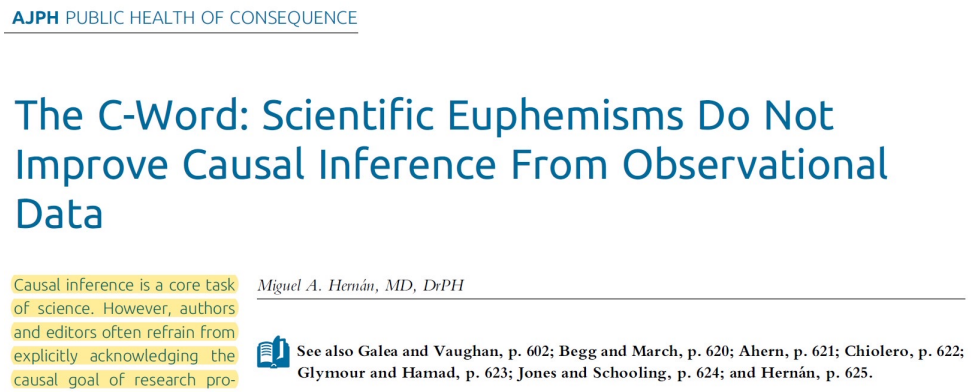
Source: Davies BMJ 1998; 316

Association vs. causation

- Our goal is to find causation
- Reality (confounding, measurement errors, study design) stops us
- We settle for association



Should we avoid causal language altogether?



AJPH PUBLIC HEALTH OF CONSEQUENCE

The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data

Causal inference is a core task of science. However, authors and editors often refrain from explicitly acknowledging the causal goal of research pro-

Miguel A. Hernán, MD, DrPH

See also Galea and Vaughan, p. 602; Begg and March, p. 620; Ahern, p. 621; Chiolero, p. 622; Glymour and Hamad, p. 623; Jones and Schooling, p. 624; and Hernán, p. 625.

Miguel Hernán argued that

- we need to use causal language to accurately describe the aims of our research
- the term “causal effect” can be appropriate
  - » in the title and Introduction section of our article when describing the aim of our research
  - » in the Methods section when describing which causal effect we are trying to estimate through an association measure
  - » in the Discussion section when providing arguments for and against the causal interpretation of our association measure.
- The only part of the article in which the term “causal effect” has no place is the Results section, which should present the findings without trying to interpret them.

Guidelines for reporting model results

Mostly dependent on the study design and nature of the data

- The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies
- The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement


THANK
YOU !